图书情报工作
第63卷 第9期 2019年5月

关键词共现网络视角下的学科基础词汇发现

于丰畅 陆伟
武汉大学信息管理学院 武汉 430072

摘要：[目的/意义]学科基础词汇是学科知识的重要基石,对于理解学科的知识体系构成、理清学科的知识脉络以及促进学科教育都有重要的意义,但长期以来其主要依赖于人工总结,目前还未实现高效地在某学科范围内自动挖掘出学科基础词汇。[方法/过程]提出一种利用关键词共现网络发现学科内较为基础的词汇的方法。该方法利用基础词汇具有相对较低的词频和在网络中具有相对较高的中心度的特性,自动从学科关键词数据集中获得该学科的基础词汇。[结果/结论]利用 ACM 中 1969 年到 2012 年的论文集的计算机领域(全数据集)、user interfaces 和 information search and retrieval 两个子主题的关键词数据集验证该方法的正确性,并且该方法能够使用较简单的步骤发现数据集中全局性的基础词汇。

关键词：共现网络 Pagerank 基础词汇
分类号：G255
DOI:10.13266/j.issn.0252-3116.2019.09.010

1 引言

基础词汇是一个学科中基础的、重要的概念和方法的载体,是理解一个学科知识的重要基石。研究如何发现某学科的基础词汇对理解该学科的知识体系构成、理清该学科的知识脉络以及促进学科教育都有着重要的意义。

对于学科基础词汇的发现,长期以来依靠人工总结,并且主要集中在初中、高中等知识体系较为简单的学科,文献中教学人员梳理初高中化学、政治等学科的基础概念并结合适当的教学法,有效地提高了学生对知识的理解程度[1-3]。在中医药学领域经过全国 10 多家中医学机构 300 多人 10 多年的努力建立了中医领域的词典性质的术语词汇库,可见人工构建某学科的词汇库在时间成本和人力成本上都面临着巨大的挑战[4]。在语言学领域,翟颖华[5]针对中国大陆的汉语《等级划分》和我国台湾地区的汉语《基础词库》的词汇研究了两岸对于用词的细微差别。以上文献从不同角度侧面印证了词汇库特别是基础词汇库具有一定的学术研究价值,并且若能在某些学科中自动挖掘学科基础词汇将会有更大的意义。

然而针对学术文本的粗粒度的知识发现已早有研究,利用共现网络或者引文网络是其中的一种重要的研究手段。将作者、论文、期刊作为网络的节点,将它们之间的共现关系或者引用关系作为边,对其构成的网络进行计量,从而得到相应的结论。P. Chen 等[6]利用物理学文献的引用关系构建网络,使用 Pagerank 算法对文献的中心度进行测量,得到了物理学领域内广为人知的基础重要文献。Y. H. Eom 等[7]利用 24 种语言的维基百科构建网络,施以 Pagerank、2DRank 和 CheiRank 算法,找到了 100 位有着重要历史地位的人物。S. Mukherjee 等[8]使用 1877 年到 2010 年的 ODI 板球比赛的历史数据,分别对球队和队长构建有向有权重的对战网络,通过对节点的出度、Pagerank 值和边的权重进行计算,得到了历史上最佳球队和最佳队长。有些学者对数字图书馆中的作者进行建模,考虑了著作的 Pagerank 值、作者信息和论文的摘要等信息,有效地推荐数字图书馆中有影响力的作者[9,10]。C. Bigonha 等[11]针对 twitter 进行了作者影响力排序的研究,该研究结合了作者在好友网络和转发网络中的位置、tweet 的极性和文本的质量,取得了较好的效果。Y. Ding 等[12]结合了 Pagerank 算法和主题模型对检索领域的学者进行了基于主题的影响力排序。Y. L. Chen[13]还首次使用 Pagerank 算法对期刊进行排序,并使用粒子群优化加入引用分析和专家意见,取得了较

作者简介：于丰畅(ORCID:0000-0002-6503-4688),博士研究生;陆伟(ORCID:0000-0002-0929-7416),教授,博士生导师,通讯作者,E-mail:weilu@whu.edu.cn。

收稿日期:2018-06-20 修回日期:2018-11-27 本文起止页码:95-100 本文责任编辑:徐健

好的效果。Z. Kozareva 等[14]运用类似的方法研究了词语粒度的挖掘，也取得了较好的结果。

可见共现网络对于挖掘关联内容具有较好的能力，笔者受到以上研究的启发，希望探究使用关键词共现网络，在计算机领域学科论文的关键词中发现该领域的基础词汇的可能性。

2　研究思路与方法

2.1　研究思路

学术论文记录了学科的发展，而论文的关键词是对论文关键内容的提炼，一般是论文中重要的概念和方法。因此，一个学科的学术论文中关键词的发展与变化在一定程度上代表了这个学科的发展情况。对于词汇粒度上的知识发现，关键词有着天然的优势，故笔者将会使用学科论文的关键词作为研究对象，从中挖掘发现一个学科的基础词汇。

文献[15－16]指出，学科主题随着时间推移会出现主题的新生、消亡、继承、分裂和合并 5 种演化形式，即基础词汇的选取应该是一个动态的过程，随着技术的不断发展，早期的尖端技术在若干年后有可能变为基础技术。特别是本文选取的计算机领域，技术快速迭代、日新月异，只针对某一时间段的基础词汇并不能正确代表该学科的发展历程。

X. Jiang 等指出观测的时间窗对于研究对象的排序结果有着较大的影响。而且，基于图的学者排名算法，在表现形式上和基于引用数量的算法虽有较大的区别，但是其结果和引用数量仍有很大的相关性[17]。故需要对两者之间的关联进行去耦合之后，才能得到较好的结果。

基于以上文献的结论，笔者将基础词汇的发现对象锁定在全局的基础词汇上，即从整个计算机科学发展的角度来发现该学科的基础词汇。并且考虑到词汇的出现频率对其中心度大小有一定的影响，笔者将采取中心度排名和频率排名的差值为指标，以此抵消过大的频率对中心度计算的影响。

在满足基础词汇是全局的前提下，必然有这些基础词汇是关于观测时间窗独立的。即：如果存在一种能够有效地发现数据集中的学科基础词汇的方法，那么从较长的观测时间窗中发现的基础词汇应该包含从较短的观测时间窗中发现的基础词汇。例如：对于时间范围是 t_0 到 t_m 的数据集，观测窗口为 t_0 到 t_m 发现的基础词集合为 F1，观测窗口为 t_0 到 t_n（$t_0 < t_n < t_m$）发现的基础键词集合为 F2，那么一定有 $F2 \subseteq F1$。即该方法满足观测窗独立性是该方法能够有效找到学科基础词的必要条件。

文献[1－3]表明，掌握学科的基础概念对掌握该学科的其他知识起着重要的作用，其背后的依据是学科内的其他概念、知识大都与基础概念有着密切的联系，基础概念在学科知识体系中起着中心作用，学科知识体系往往是从基础概念出发的网状结构。即在后序的关键词共现网络中，该方法发现的基础词汇是网络中中心度较高的词，这是该方法能够有效找到学科基础词汇的另一个必要条件。

可以从该命题的否命题定性地证明该命题是一个必要条件。该命题的否命题为：基础词汇不是所在关键词共现网络中中心度较高的关键词。根据后文 2.2 节对关键词共现网络的中心度的计算方法的特点，若一个节点在该网络中中心度不够高，其原因有且仅有两种：①该节点相连的节点中心度均较低，即该关键词较少与重要的关键词共现；②该节点相连的其他节点有较高的中心度，但同时也有较高的度数，即该关键词是一个“百搭”的词汇，其作为描述该学科的词汇时精准性不高。两种原因都不符合前文中对学科基础词汇的定义，即该假设的否命题不成立，故该命题也是发现学科基础词汇的一个必要条件。

2.2　Pagerank 算法

本研究选取计算机文献的关键词作为网络的节点，同一篇文献中同时出现的关键词的共现关系作为网络的边。利用 Pagerank 算法计算网络中节点的中心度，Pagerank 算法的核心思想建立在互联网随机冲浪者模型之上[18]，算法表达如公式 1 所示：

$$G_i = \frac{\lambda}{N} + (1 - \lambda) \sum_{j\,nm\,i} \frac{G_j}{K_j} \qquad \text{公式(1)}$$

一个网页的 Pagerank 值由两部分相加而成。加号右边的一项表示和网页 i 相连接的所有网页对网页 i 的贡献，累加符号表示和节点 i 邻近的所有节点 j。K_j 表示网页 j 的度数，即网页 j 对与其相连的网页在 Pagerank 值上有均等的贡献，均为其 Pagerank 值的 K_j 分之一。加号左边的一项表示，由网络上的任意网页跳转到该网页 i 上时贡献的 Pagerank 值，其中 N 为网络中所有网页的个数，λ 称作阻尼系数，对于公式(1)，G_i 将会随着 λ 的增大逐渐趋于 $\frac{\lambda}{N}$，这就意味着过大的 λ 将导致网络中的所有节点的 Pagerank 值趋于一致，使得网络中节点的区分度降低。相反，如果 λ 逐渐减小，G_i 会在更大程度上受到节点 i 周围的节点影响，换句

话说如果 i 节点周围节点的 Pagerank 值越大,i 节点的 Pagerank 值将会进一步增大,有利于区分网络中不同重要性的节点。故本研究用将沿用 L. Page 和 S. Brin 在最初论文[18]中使用的 λ = 0.15 的参数设置。

Pagerank 算法有如下 3 点特性:①节点 i 分别与节点 j 和 k 相连并且 j 和 k 有相同的度时,若节点 j 的 Pagerank 值大于节点 k,那么节点 j 对于节点 i 的 Pagerank 值贡献更大;②与节点 i 相连的具有相同 Pagerank 值节点中度数较少的节点对 i 的 Pagerank 值贡献较大;③当节点 i 与众多节点相连时,节点 i 的 Pagerank 值也会较大。

3 数据集

本研究使用了国际计算机学会(Association for Computing Machinery)的全文数据集。从中挑选了 1951－2012 年的 215 710 篇论文,实验前对数据进行清理,去除了数据集中未包含关键词的论文,保留了 110 363 篇包含关键词、年限分布在 1969－2012 年之间的论文。数据集中包括 364 个子主题分类,按照主题下论文数量进行排序,包含论文数量最多的两个子主题分别为 user interfaces 和 information search and retrieval。笔者将对整个计算机学科(全数据集)、user interfaces 主题和 information search and retrieval 主题分别进行实验,以此验证对基础词汇发现的方法的正确性。

图 1 显示了数据集中论文的分布情况。由于数据集是在 2012 年中收集的,故除 2012 年以外论文的数量和包含关键词的论文数量均呈现逐年上升的趋势。图 2 为含关键词的论文在当年论文中的占比情况,其中纵坐标为以 10 为底数的对数,该图从 1969 年开始统计的原因是,在此之前的数据集中的论文均不包含关键词。从 1990 年开始,包含关键词的论文占比逐年增加,并且到 2011 年包含关键词的论文已经接近当年论文总数的 90%,考虑到含关键词率较低的年份论文总数也相对较少,故本文中使用的数据集能够较好地代表计算机领域的研究状况。

关键词作为论文作者对于其科研成果的归纳和提炼,包含了作者选取的核心主题和重要方法,故关键词在一定程度上反映了论文的核心内容。本研究选取关键词作为反应学科基础知识的原因也在于此。所以,数据集中关键词若是由作者本人选取,将对本研究的准确性提供更大的帮助。笔者按年份随机抽取了数据集中的 30 篇论文,人工比对了数据集中的关键词和原文中作者提供的关键词,抽样的论文的关键词和数据集中的关键词均能吻合。

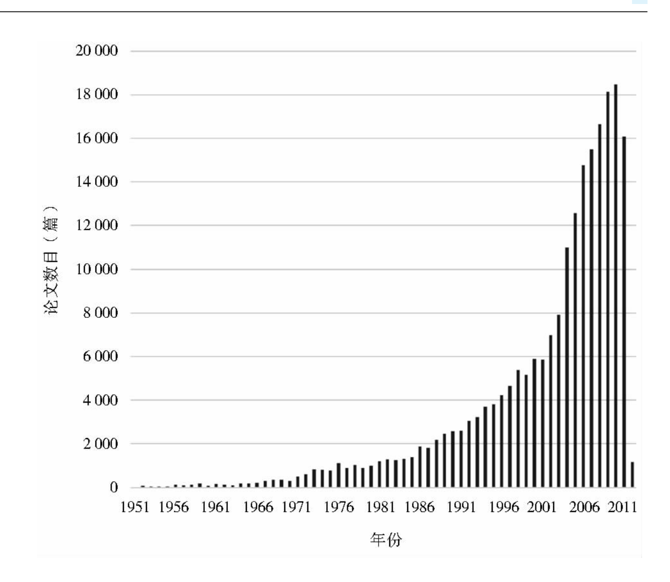


图 1 数据集中论文关于时间的分布

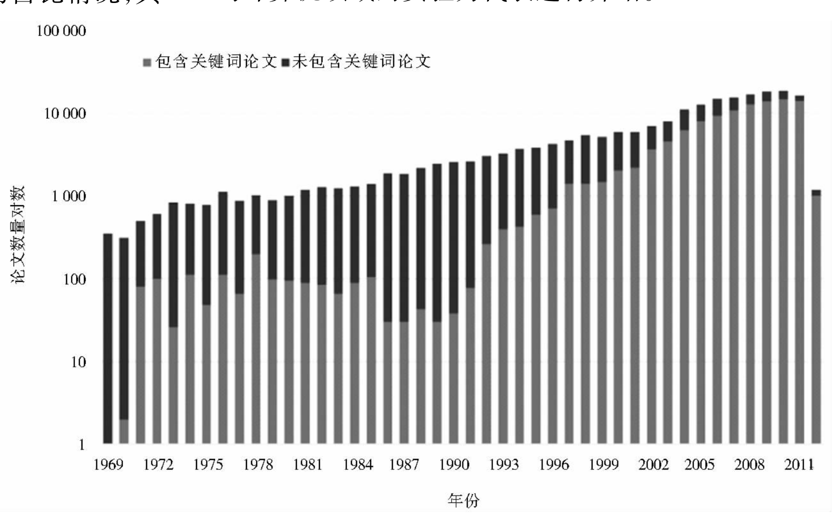


图 2 数据集中包含关键词的论文和未包含关键词的论文的分布

4 实验

4.1 实验设计

为了验证该方法是否能有效地发掘计算机领域内的基础词汇,实验将分为 3 组,分别针对计算机领域(整个数据集)、数据集中的 user interfaces 和 information search and retrieval 两个子主题进行实验。下面以均计算机领域的实验为代表进行介绍。

实验流程如下：

（1）按照时间窗口将数据集分段；

（2）对每个窗口的关键词构建共现网络；

（3）计算每个网络的节点 Pagerank 值和 TF 值，找到两值排名差异大的关键词；

（4）对步骤 3 中各个时间窗口的结果取交集。

4.2　观测时间窗设置

按照研究思路中的第一个必要条件，需要将数据集分为 T 个时间窗，本文将数据集中的论文按照时间顺序分为 5 个存在重叠的时间窗口（即 T = 5），使得时间窗内的论文数量按照等差进行分布。数据集共有含关键词的论文 110 363 篇，5 个观测窗对应的论文数量分别为 22 073 篇、44 146 篇、66 219 篇、88 292 篇、110 363篇，对应的观测时间为 1969 – 2004 年、1969 – 2007 年、1969 – 2008 年、1969 – 2010 年、1969 – 2012 年。笔者如此设计时间窗口，是为了使每个观测时间窗均能保证从数据集的时间开端开始研究学科的发展。

4.3　网络构建

根据以上 5 个观测窗口，利用 Python 的 Networks 工具包分别对 5 个窗口内的关键词构建成 5 个双向有权重的共词网络，节点为关键词，节点之间的边代表两个关键词同时作为某一篇论文的关键词，每条边的权重均为 1。特别地，对于第 5 个时间窗所构成的网络，其节点的集合为 V，由 V 构成的图为 G。

4.4　网络节点计算

对于以上 5 个共词网络，利用 2.2 节中介绍的 Pagerank 算法和参数设置，对这 5 个共词网络分别计算每个节点的中心度（Pagerank 值）。同时分别计算每个关键词在时间窗内出现的次数，即词频（TF）。分别得到每个时间窗内的关键词按照 Pagerank 的排名（GRank）和按照 TF 的排名（tfRank）。图 3 所示为第 5 个时间窗内，关键词的出现次数（TF）与平均中心度（Pagerank）之间的关系，其中每一个点代表一个关键词，虚线为利用图中所有点拟合出的一根经过原点的直线。从图 4 中可以看出，绝大多数的关键词的频率与 Pagerank 值服从正比例关系，这也符合文献[17]中研究的结论，即更大的出现频率一般会伴随更大的 Pagerank 值，也即高 tfRank 的关键词一般会对应高 GRank 的排名。

4.5　学科基础词汇发现

根据研究思路中的第二个必要条件，基础词汇为 Pagerank 值较高的关键词，故需要取出每个时间窗中

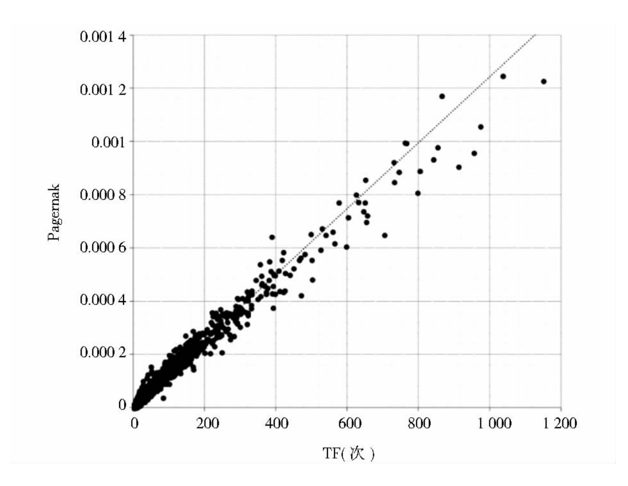


图 3　关键词的词频和 Pagerank 值的对应关系

GRank 按从小到大排序中前 Topg% 对应的关键词。

为了得到 4.1 节步骤（3）中 Pagerank 值的排名和 TF 值的排名差异大的关键词，还需要将 GRank 前 Topg% 对应关键词的 GRank 与 tfRank 做差。所得的结果如果为负，代表该关键词拥有较低的 TF 和较高的 Pagerank 值。将做差之后的结果按照差值从小到大排序，取出前 Topt% 的结果对应的关键词，作为一个观测时间窗内的候选基础词结果。

需要说明的是，根据 2.2 节的公式（1），Pagerank 算法得出的结果是一个没有量纲的值，而 TF 值统计的是某关键词在数据集中出现的次数，其量纲为"次"。两者的量纲不同，不能直接做减法运算，故使用两个排名做减法运算来表征两者的差异大小。

参数 Topg 决定了被选择的关键词的观测时间窗内的重要程度，而参数 Topt 决定前叙关键词的两个排名的差异程度，并且这两个参数也控制着最终找到的基础词汇的数量。考虑到某一学科的基础词汇总量相对有限，经过多次实验，本文中，对于计算机学科选取 T = 5，Topg = 3，Topt = 33，对于 user interfaces 和 information search and retrieval 两个子主题选取 T = 4，Topg = 10，Topt = 25。

根据研究思路中的第一个必要条件，全局的基础词汇满足时间窗独立性，故将得到的 5 组结果取交集，所得到的结果即是本方法发现的计算机学科的基础词汇。

4.6　结果验证

笔者尚未发现现存类似的计算机学科基础词汇表可供对比，为了验证结果的正确性，笔者采用人工检验的方式对 3 次实验的结果进行评测，将这 3 次实验结果分别交由 3 个领域的各 1 名副教授或博士后进行人工检验。

在检验之前笔者给评测者详细地解释本文中基础词汇代表某学科的基础的、重要的概念和方法这一标准。在测试中,评测者将勾选其认为满足这样的定义的词汇。

表 1 为针对 3 次实验的人工测评结果,人工评测的正确率的计算方法是评测者勾选的基础词汇的个数除以提供给评测者的词汇总个数,3 次实验经过前述方法计算得出的基础词汇数量分别为,计算机领域 232 个,information search and retrieval 子主题 110 个,user interfaces 子主题 153 个。

表 1　针对 3 次实验的人工测评结果

实验对象	计算机领域	information search and retrieval 子主题	user interfaces 子主题
人工评测正确率	91.81%	84.55%	86.27%

从测试结果中可以看出,利用本文提出的方法发现的学科基础词汇准确率较高,并且在计算机领域的人工评测准确率最高。笔者推测的原因是计算机领域涉及范围相对最广,存在的基础词汇数量相对更多,而提供给评测者的待检验基础词汇数量又相对有限,故计算机领域的基础词汇的的准确率相对较高。

5　分析与讨论

观察实验所得到的结果,不难发现结果中包括了 data structure、network topology、microprocessors、time complexity、parallel algorithm、web site、program debugging 等典型的计算机科学的基础词汇。比如数据结构(data structure)是计算机中储存、组织数据的方式,是计算机程序设计中的基础环节。又如微处理器(microprocessors)或者称为中央处理器,是计算机硬件中最为核心的一部分,执行电路控制和逻辑运算等重要功能。诸如数据挖掘、虚拟现实、机器学习、云计算等计算机领域近 10 年来热门的前沿词汇均没有出现在本方法挖掘的结果之中。

表 2 为计算机领域的前沿关键词(前 5 行)与本方法中挖掘得到的基础词汇(后 5 行)的对比情况。从表 2 中可以看出前 5 行的关键词同时具有较高的 Pagerank 值与较高的词频,并且两者的排名差异较小。而基础词汇的词频相对低,Pagerank 值相对较高。该方法本质上是在数据集中寻找拥有较低词频且中心度较高的关键词。

图 4 展示了基础词汇在网络拓扑结构中的特殊之处,以表 1 中的关键词为根,图 G 中与其连接的所有关键词为叶子所构成的树。其中每一棵树都是 G 的一个

表 2　前沿关键词和基础词汇的对比

关键词类型	关键词	TF	GRank	tfRank	GRank-tfRank
前沿关键词	Data mining	729	10	15	－5
	Wireless sensor networks	731	15	14	1
	virtual reality	630	18	22	－4
	machine learning	648	19	20	－1
	cloud computing	358	55	64	－9
基础关键词	data structure	36	944	1 504	－560
	network topology	38	1 079	1 453	－374
	microprocessors	26	1 632	2 252	－620
	time complexity	18	2 476	3 166	－690
	parallel algorithm	35	972	1 562	－590

子图,所有的根的节点大小均设置为 1,即所有叶子节点的大小为其关键词的 Pagerank 值对根节点归一化之后的结果。

从图论的角度观察,前沿关键词拥有高 GRank 排名的关键词,同时也具有较高的度数,且所有叶子节点的 Pagerank 值都相对较低。而学科基础词,虽然度数相对较低但所有叶子节点的 Pagerank 值都相对较高。

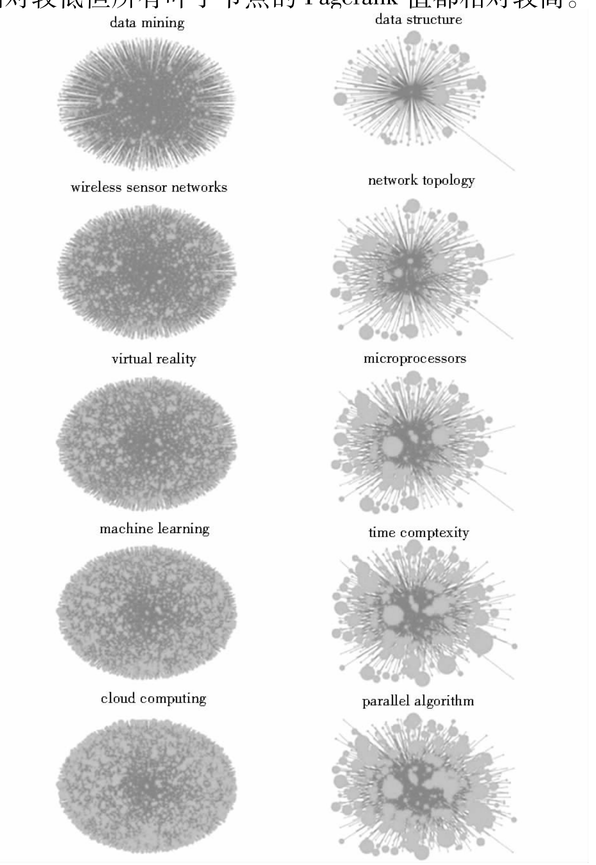


图 4　前沿关键词和基础关键词在共现网络中的对比

6 结语

笔者设计了一套发现学科基础词汇的方法,并利用 ACM 数据集以及该数据集中两个子主题验证了这套方法的有效性。该方法具有简单有效的特点,可以快速地找到领域内的基础词汇。但是,本研究也存在着一些局限,例如无法对基础词汇进行排序,后续的研究工作应改进度量方法,进一步对学科词汇的基础性进行计算。

参考文献:

[1] 温海港. 探讨高中阶段化学基础概念的教学[J]. 考试周刊,2015(56):131.
[2] 马晓敏. 初中化学基础概念知识网络构建——思维导图在教学和学习中的使用[J]. 赤子(上中旬),2015(6):317.
[3] 余静. 向基础概念要成绩——论高中思想政治概念教学[J]. 当代教研论丛,2015(2):116,118.
[4] 贾李蓉,李海燕,于彤,等. 中医药学语言系统基础词库分析[J]. 中国数字医学,2014,9(2):66-67.
[5] 翟颖华. 新一代两岸初级汉语词表比较及引发的思考[J]. 华文教学与研究,2015(2):43-52.
[6] CHEN P, XIE H, MASLOV S, et al. Finding scientific gems with Google's pagerank algorithm[J]. Journal of informetrics, 2007, 1(1): 8-15.
[7] EOM Y H, ARAGÓN P, LANIADO D, et al. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions[J]. PLOS ONE, 2015, 10(3): e0114825.
[8] MUKHERJEE S. Identifying the greatest team and captain - a complex network approach to cricket matches[J]. Physica a: statistical mechanics and its applications, 2012, 391(23): 6066-6076.
[9] MIMNO D, MCCALLUM A. Mining a digital library for influential authors[C]//Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries. New York:ACM, 2007: 105-106.
[10] LIN L, XU Z, DING Y, et al. Finding topic-level experts in scholarly networks[J]. Scientometrics, 2013, 97(3): 797-819.
[11] BIGONHA C, CARDOSO T N C, MORO M M, et al. Sentiment-based influence detection on Twitter[J]. Journal of the brazilian computer society, 2012, 18(3): 169-183.
[12] DING Y. Topic-based PageRank on author co-citation networks[J]. Journal of the Association for Information Science and Technology, 2011, 62(3): 449-466.
[13] CHEN Y L, CHEN X H. An evolutionary pagerank approach for journal ranking with expert judgements[J]. Journal of information science, 2011, 37(3): 254-272.
[14] KOZAREVA Z, HOVY E. Insights from network structure for text mining[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2011: 1616-1625.
[15] 曲佳彬, 欧石燕. 基于主题过滤与主题关联的学科主题演化分析[J]. 数据分析与知识发现, 2018, 2(1): 64-75.
[16] 唐果媛. 基于共词分析法的学科主题演化研究方法的构建[J]. 图书情报工作, 2017, 61(23): 100-107.
[17] JIANG X, SUN X, ZHUGE H. Graph-based algorithms for ranking researchers: not all swans are white! [J]. Scientometrics, 2013, 96(3): 743-759.
[18] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the Web[R]. Stanford InfoLab, 1999.

作者贡献说明:
于丰畅:数据处理分析,论文撰写;
陆伟:理论指导和思路设计。

The Discovery of Subject Basic Vocabulary from the Perspective of Keyword Co-occurrence Network

Yu Fengchang Lu Wei
School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Subject basic vocabulary is an important cornerstone of subject knowledge. It is of great significance to understand the composition of the knowledge system of discipline, to clarify the knowledge context of discipline and to promote discipline education. However, for a long time, it mainly relies on manual summarization and cannot be automatically mined within a certain discipline. [**Method/process**] This paper proposes a method to use the keyword co-occurrence network to discover basic vocabularies within the discipline. This method takes advantage of the relatively low word frequency of the basic vocabulary and the relatively high degree of centrality in the network, and automatically obtains the subject basic vocabulary from the subject keyword dataset. [**Result/conclusion**] The validity of this method is verified by using the keyword datasets in the fields of computer(full dataset), user interfaces and information search and retrieval from ACM's 1969-2012 theses. Moreover, this method can use simpler steps to discover the global basic vocabulary in the data set.

Keywords: co-occurrence network pagerank subject basic vocabulary